PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2001-318945
(43)Date of publication of application : 16.11.2001

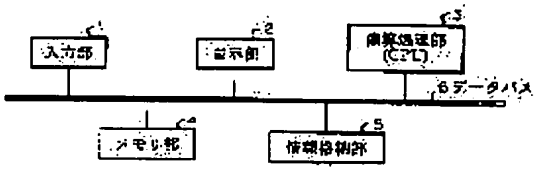
(51)Int.Cl. G06F 17/30
G06F 17/21

(21)Application number : 2000-139767
(22)Date of filing : 12.05.2000

(71)Applicant : RICOH CO LTD
(72)Inventor : NARITA MASUMI
OGAWA YASUTSUGU

(54) DOCUMENT RETRIEVING DEVICE AND ITS METHOD

(57)Abstract:
PROBLEM TO BE SOLVED: To obtain a highly accurate retrieval result by extracting a word and a noun phrase suitable for retrieving an inputted retrieving request and setting up a suitable retrieving condition.
SOLUTION: A language analysis means analyzes the language of a retrieving request inputted by a retrieving request input means 11 and extracts words to constitute the elements of a retrieving condition and a noun phrase composed of a plurality of words. A retrieving condition generation means 13 combines the extracted words and noun phrase by a prescribed operator and applying prescribed weight to each of the words and the noun phrase to generate a retrieving condition. A document retrieving means 14 extracts a document which agrees with the retrieving condition generated by the means 13 from a document database 15, and a retrieved result display means 16 displays the extracted result.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

引用文献8

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号
特開2001－318945
（Ｐ２００１－３１８９４５Ａ）
(43)公開日　平成13年11月16日(2001.11.16)

(51)Int.Cl.[7]	識別記号	ＦＩ		テーマコード（参考）
G06F 17/30	330	G06F 17/30	330	C 5B009
	170		170	A 5B075
	210		210	B
17/21	590	17/21	590	E

審査請求　未請求　請求項の数５　ＯＬ　（全８頁）

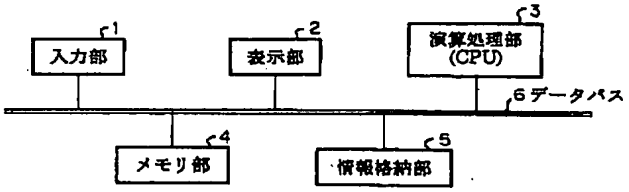
(21)出願番号　特願2000－139767(Ｐ2000－139767)
(22)出願日　平成12年５月12日(2000.5.12)

(71)出願人　000006747
株式会社リコー
東京都大田区中馬込１丁目３番６号
(72)発明者　成田　真澄
東京都大田区中馬込１丁目３番６号　株式会社リコー内
(72)発明者　小川　泰嗣
東京都大田区中馬込１丁目３番６号　株式会社リコー内
(74)代理人　100079843
弁理士　高野　明近　（外２名）

Ｆターム(参考)　5B009 MB16 VA02
5B075 ND03 NK32 NK33 PP02 PP03
PP26 PQ02 UU06

(54)【発明の名称】文書検索装置及び文書検索方法

(57)【要約】
【課題】　入力された検索要求を検索に適切な単語と名詞句を抽出して、適切な検索条件を設定することにより、精度の高い検索結果を得る。
【解決手段】　検索要求入力手段１１より入力された検索要求を言語解析手段により言語解析して検索条件の要素となる単語及び複数の単語より構成された名詞句を抽出する。抽出した単語及び名詞句を検索条件生成手段１３により、所定の演算子により結合し、かつ、単語及び名詞句の各々に所定の重み付けを施して検索条件を生成する。検索条件生成手段１３により生成した検索条件に合致した文書を文書検索手段１４により文書データベース１５中の検索対象文書から抽出し、抽出した結果を検索結果表示手段１６にて表示する。



1

【特許請求の範囲】

【請求項１】　検索対象である複数の文書から検索条件に合致した文書を抽出する文書検索装置であって、入力された検索要求を言語解析して検索条件の要素となる単語及び複数の単語より構成された名詞句を抽出する言語解析手段と、抽出した単語及び名詞句を所定の演算子により結合し、かつ、単語及び名詞句の各々に所定の重み付けを施して検索条件を生成する検索条件生成手段と、該検索条件生成手段により生成した検索条件に合致した文書を検索対象文書から抽出する文書検索手段とを有することを特徴とする文書検索装置。

【請求項２】　請求項１記載の文書検索装置において、前記言語解析手段は、検索条件の要素となる単語を抽出する際には予め作成した不要語リストを使用して検索要求から不要な単語を削除し、名詞句を抽出する際には不要語リストの使用条件を緩めることを特徴とする文書検索装置。

【請求項３】　請求項１または請求項２記載の文書検索装置において、前記検索条件生成手段は、前記言語解析手段によって抽出された名詞句が３つ以上の単語で構成される場合には、前記言語解析手段によって名詞句の各構成単語に付与された品詞情報に基づいて２単語からなる単語ペアを選択して検索条件を生成するようにしたことを特徴とする文書検索装置。

【請求項４】　請求項１または請求項２記載の文書検索装置において、前記検索条件生成手段は、前記言語解析手段によって抽出された名詞句の中に他の句構造を持つ要素が含まれている場合には、名詞句全体の句構造情報を基にして主要な語句を抽出して検索条件を生成するようにしたことを特徴とする文書検索装置。

【請求項５】　検索対象である複数の文書から検索条件に合致した文書を抽出する文書検索方法であって、入力された検索要求を言語解析して検索条件の要素となる単語及び複数の単語より構成された名詞句を抽出し、抽出した単語及び名詞句を所定の演算子により結合し、かつ、単語及び名詞句の各々に所定の重み付けを施して検索条件を生成し、生成した検索条件に合致した文書を検索対象文書から抽出する各段階を有することを特徴とする文書検索方法。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、文書検索装置及び文書検索方法に係り、特に、電子化された文書情報から検索要求に合致する文書を検索するための文書検索装置及び文書検索方法に関する。

【０００２】

【従来の技術】複数の文書情報を格納した文書データベースから特定の文書を検索するために文書検索装置が用いられる。このような文書検索装置は、入力された検索要求に合致する文書情報を文書データベースから抽出するものである。一般的に、入力された検索要求の内容をそのまま検索条件として使用することはできず、実際の検索に使用される検索条件は文書検索装置により生成される場合が多い。すなわち、ユーザが入力する検索要求は、検索に不要な語句を含んでいる場合が多いので、入力された検索要求を言語解析して検索に不要な語句を除去することにより、検索に必要な語句のみを抽出する処理が行われる。

2

【０００３】特開平６－７５９９６号公報では、与えられた検索要求を言語解析することにより検索条件を生成する方法を開示している。この方法では、入力された検索要求文に対して形態素解析を適用して検索要求文中の各々の単語を識別し、識別した単語を活用形へ展開したり、複合語を分解したりして、検索要求の同義表現を生成する。そして、検索要求語およびその同義表現と文書データベースとを照合して文書検索を行い、ユーザの検索意図に合致した文書を検索する。

【０００４】

【発明が解決しようとする課題】上記特開平６－７５９９６号公報に記載された発明で、

（１）分解の対象となる複合語、あるいは複数の単語よりなる句として、２単語から構成されるものに限られており、３単語以上から構成される複合語あるいは句に対応できない。

（２）検索語とその同義表現に対して同じ重み付けで検索条件が設定されており、精度の高い検索結果を得ることができない。

（３）検索要求語から同義表現を生成する際に、助動詞や助詞といった機能語を検索に不要な語句として除去しているが、複合語を構成している内容語について複合語が分解されたときに除去の対象とすべきものがあるかどうかという考慮がなされていない。つまり、検索に不要な語句として内容語に相当するものを設定しているかどうかが不明であり、複合語が分解されて単一の単語からなる検索要求が生成されたとき、非常に使用頻度の高い内容語が検索に使用されて検索精度の低下を招く恐れがある。

【０００５】本発明は、上述の問題点に鑑みなされたものであり、自然言語で入力された検索要求を言語解析により検索に適切な単語と名詞句を抽出し、抽出された単語及び名詞句に基づいて適切な検索条件を設定することにより、精度の高い検索結果を得ることを目的とする。

【０００６】

【課題を解決するための手段】上述の目的を達成するために、請求項１記載の発明は、検索対象である複数の文書から検索条件に合致した文書を抽出する文書検索装置であって、入力された検索要求を言語解析して検索条件の要素となる単語及び複数の単語より構成された名詞句を抽出する言語解析手段と、抽出した単語及び名詞句を所定の演算子により結合し、かつ、単語及び名詞句の各

々に所定の重み付けを施して検索条件を生成する検索条件生成手段と、該検索条件生成手段により生成した検索条件に合致した文書を検索対象文書から抽出する文書検索手段とを有するもので、単語と名詞句に適切な重み付けを施して演算子により結合して検索条件が生成される。これにより、名詞句単位での検索条件と単語単位での検索条件とを合わせて検索条件を生成することができるので検索漏れを低減し、検索精度を高めることができる。

【０００７】請求項２記載の発明は、請求項１記載の文書検索装置であって、前記言語解析手段は、検索条件の要素となる単語を抽出する際には予め作成した不要語リストを使用して検索要求から不要な単語を削除するが、名詞句を抽出する際にはこの不要語リストの使用条件を緩めるもので、不要語としてリストアップされた内容語でも名詞句を構成する場合には名詞句としての意味を表す上で重要なものとなるため、名詞句を抽出する際には不要語リストにある内容語は削除対象としないようにした。これにより、名詞句の構成単語に不要語リストにある内容語が含まれていても名詞句としての検索条件生成の対象となるので検索漏れを低減し、検索精度を高めることができる。

【０００８】請求項３記載の発明は、請求項１または請求項２記載の文書検索装置であって、前記検索条件生成手段は、前記言語解析手段によって抽出された名詞句が３つ以上の単語で構成される場合に、前記言語解析手段によって当該名詞句の各構成単語に付与された品詞情報に基づいて２単語からなる単語ペアを選択するもので、３つ以上の単語で構成される名詞句での検索条件は２つの単語で構成される名詞句より一層厳しい検索条件となるため、２つの単語で構成される単語ペアに分割して検索条件を生成することにより、検索条件の範囲を広げることができる。このとき、言語解析手段によって付与された品詞情報に基づいて検索に使用する単語ペアを選択するので、検索条件として不適当な単語ペアを予め除外することができる。

【０００９】請求項４記載の発明は、請求項１または請求項２記載の文書検索装置であって、前記検索条件生成手段は、前記言語解析手段によって抽出された名詞句の中に他の句構造を持つ要素が含まれている場合には、名詞句全体の句構造情報を基にして主要な語句を抽出して検索条件を生成するもので、これにより、前置詞句が埋め込まれているような深い内部構造を有する名詞句からも適切な検索条件を生成することができ、検索条件の範囲を広げることができる。

【００１０】請求項５記載の発明は、検索対象である複数の文書から検索条件に合致した文書を抽出する文書検索方法であって、入力された検索要求を言語解析して検索条件の要素となる単語及び複数の単語より構成された名詞句を抽出し、抽出した単語及び名詞句を所定の演算子により結合し、かつ、単語及び名詞句の各々に所定の重み付けを施して検索条件を生成し、生成した検索条件に合致した文書を検索対象文書から抽出する各段階を有するものである。

【００１１】

【発明の実施の形態】図１は、本発明の第１の実施の形態による文書検索装置のブロック図で、本発明の実施の形態による文書検索装置は、入力部１と、表示部２と、中央演算装置（ＣＰＵ）を含む演算処理部３と、メモリ部４と、情報格納部５と、これらを接続するデータパス６よりなる。入力部１は、キーボード、マウス、タッチパネル等により構成され、ユーザが文書検索装置に情報を入力するために使用される。表示部２は、ＣＲＴディスプレイあるいは液晶ディスプレイ等よりなり、文書検索装置により得られた情報をユーザに対して表示したり、入力部１から入力された情報を表示する。演算処理部３は、所定のプログラムに基づいて文書検索処理を行う。メモリ部４は、演算処理部３が実行するプログラムを格納するＲＯＭと演算処理部が動作するときに必要な情報を一時的に格納するＲＡＭとにより構成される。情報格納部５は、ハードディスク装置等の比較的大容量の記憶装置よりなり、検索対象となる文書群が登録された文書データベースやプログラムを格納する。

【００１２】図２は、図１に示した文書検索装置の機能ブロック図で、図２における矢印は文書検索装置内の処理の流れを示している。図２において、検索要求入力手段１１は、ユーザが検索したい文書の内容を記述した自然言語を入力する機能を有するものであり、入力部１の機能に相当する。ここで、自然言語とは、例えば日本語、英語、独語、仏語等のような言語を意味し、検索対象となる文書も自然言語で表記されたものとする。ここで、ユーザが検索したい文書の内容を記述した情報を検索要求と称する。検索入力手段１１により入力された検索要求は言語解析手段１２に供給される。

【００１３】言語解析手段１２は、演算処理部（ＣＰＵ）３が所定のプログラムを実行することにより達成される。すなわち、言語解析手段１２は、検索要求を形態素解析して検索要求中の各々の単語を認識し、認識した単語の中から検索条件に適当な単語を抽出する。また、言語解析手段１２は、単語の品詞情報を基にした句分割規則を使用して、名詞句としてまとめられる単語群を抽出する。この処理は、名詞句分割と称される処理であり、言語解析の分野では周知の処理であるので、その説明は省略する。また、形態素解析処理も、言語解析の分野では周知の処理であり、その説明は省略する。

【００１４】検索条件生成手段１３は、言語解析手段１２による処理結果を受け取り、抽出された単語及び名詞句を適切な演算子で結合して検索条件を生成する。検索条件生成手段１３は、演算処理部（ＣＰＵ）３が所定のプログラムを実行することにより達成される。検索条件

生成手段１３により生成された検索条件は、文書検索手段１４に供給される。

【００１５】文書検索手段１４は、文書データベース１５に登録された文書情報を検索して、供給された検索条件に合致する文書情報を抽出する。文書検索手段１４は、演算処理部（ＣＰＵ）３が所定のプログラムを実行することにより達成される。文書検索手段１４により抽出された文書情報は、検索結果表示手段１６に供給される。

【００１６】検索結果表示手段１６は、表示部２の機能に相当し、検索結果として抽出された文書情報を表示する。これにより、検索要求を入力したユーザは検索結果を表示画面上で確認することができる。また、表示部２にプリンタを設けることにより、検索結果を印刷してもよい。

【００１７】次に、上述の言語解析手段１２の処理結果について説明する。言語解析手段１２の処理は、従来の言語解析手法を用いて行われるため、処理結果についてのみ説明する。ユーザは、入力部１（キーボード）を操作して検索要求を入力する。通常、検索要求はユーザの検索意図を表した一つの文章として入力される。このユーザが入力する文章を検索要求文と称することとする。

【００１８】図３は、検索要求入力手段１１により検索要求文を入力したときの入力画面の一例である。検索要求文として“Find information on security measures which will gointo effect in airports.”という英語の文章が入力されている。入力画面中の「初期件数」は検索結果として検索条件に合致する文章を３０件表示することを指定している。検索要求文には、例えば、冠詞、前置詞、接続詞といった検索に必要のない単語が含まれている。従って、言語解析手段１２は、検索要求文から検索に不要な語句を除去して検索に必要な語句のみを抽出する。不要な単語の除去は、予め作成しておいた不要語リストを参照しながら行われる。不要語リストには、冠詞、前置詞、接続詞等の機能語や、ユーザの検索意図に関連しないと考えられる内容語が登録されている。すなわち、言語解析手段１２は、検索要求文の各々の単語を不要語リストと照合し、不要語リストに登録されている単語を除去することにより、検索に使用する単語を抽出する。

【００１９】一方、言語解析手段１２による名詞句分割処理の結果として得られた名詞句については、名詞句を構成している単語と不要語リストの照合が行われ、不要語リストに登録されている機能語のみが名詞句中の除去対象の単語となる。これは、例えば、“What types of cases were heard by the World Court”という英語の検索要求文に対して、言語解析手段１２によって同定された“the World Court”という名詞句から、不要語リストにある機能語“the”が除去されることを指す。

【００２０】ここで、ユーザが検索要求文として、図３に示すように、“Find informationon security measures which will go into effect in airports.”という英文を入力したものとする。この検索要求文に対する言語解析手段１２の処理結果を以下に説明する。入力された検索要求文は１２個の単語で構成されているが、予め作成しておいた不要語リストと照合することにより適切な単語のみが抽出される。本実施の形態では、不要語リストに“find”，“information”，“on”，“measure”，“which”，“will”，“go”，“into”，“effect”，“in”が登録されているものとする。

【００２１】単語単位の検索語は検索要求文からこれらの単語をすべて除去することによって、また、名詞句単位の検索語は言語解析手段１２によって名詞句として同定された“security measures”において不要語リストに“measure”が含まれていても内容語であるために除去の対象とならないことにより、以下のものが抽出される。

抽出された単語（２単語）：security, airports

抽出された名詞句（１名詞句）：security measures

【００２２】この抽出された２単語及び１名詞句が、検索条件を生成する要素となる。検索条件生成手段１３は、抽出された単語と名詞句に対して適当な重み付けを施し、演算子で結合することにより検索条件を生成する。演算子としては、ＡＮＤ，ＯＲのような論理演算子が使用される。また、近傍演算子としてＷＩＮＤＯＷが使用され、重み付け演算子としてＳＣＡＬＥが使用される。

【００２３】演算子ＡＮＤは、検索される文書中にこの演算子で結合された単語の全てが含まれる場合にその文書を検索結果として抽出することを指定するための演算子である。演算子ＯＲは、検索される文書中にこの演算子で結合された単語のいずれか一つが含まれる場合にその文書を検索結果として抽出することを指定するための演算子である。

【００２４】また、演算子ＷＩＮＤＯＷは、この演算子で結合される２つの単語の間の距離と語順を指定するための演算子であり、例えば、#window[1, 1, o]といった形式で表記される。括弧内の最初の数字と２番目の数字により単語の出現する範囲が規定され、３番目の文字は２つの単語の語順を表わしており、′o′は表記されたとおりの順序で２つの単語が出現することを指定している。すなわち、上記の例では２つの単語が表記された順番で隣接して出現することが指定される。

【００２５】また、演算子ＳＣＡＬＥは単語単位での検索条件と名詞句単位での検索条件とで重み付けの調整を行うための演算子である。例えば、#scale[0.5]というように表記した場合、これに続く検索条件の重み付けを０．５とすることを表わす。本実施の形態では、単語と名詞句とに異なる重み付けを施すことにより、検索結果の精度を向上させている。本発明者は、様々な試行の結

果、名詞句単位の検索条件に対する重み付けを単語単位の検索条件に対する重み付けより小さくすることにより、検索精度が向上することを見出した。本実施の形態では、各単語単位の検索条件に対する重み付けを１とし、名詞句単位の検索条件に対する重み付けを０．５としている。

【００２６】上述の演算子を使用して、本実施の形態において上述の検索要求文から生成した検索条件は以下のようになる。#or (security, airports, #scale [0.5] (#window [1, 1, o] (security, measures)))検索条件生成手段１３により上記のような検索条件が生成されると、文書検索手段１４は文書データベース１５に登録された文書のうち検索条件に合致する文書を抽出する。このとき、検索条件に対して重み付けを考慮して得られた各々の文書のスコアを比較し、スコアの高い文書を検索条件に合致した文書として抽出する。このような文書検索処理として周知の文書検索処理を用いており、その説明は省略する。

【００２７】文書検索処理が終了すると、検索結果表示手段１６は、図４に示すように、検索結果としてスコアの高い文書から順番に画面に表示する。ここで、初期件数として３０件を表示することが指定されているため、スコアの高い順から３０件の文書を画面に表示する。図４の画面において、画面をスクロールすることにより、検索結果として抽出された３０件の文書を閲覧することができる。

【００２８】次に、本発明の他の実施形態、すなわち、言語解析手段によって抽出された名詞句が３つ以上の単語で構成される場合について説明する。全体の機能構成は図２と同じであり、相違点は検索条件生成手段１３において、言語解析手段１２で抽出された名詞句が３つ以上の単語で構成される場合に、言語解析手段１２によって当該名詞句の各構成単語に付与された品詞情報に基づいて２単語からなる単語ペアを選択することにある。これは、検索要求文で用いる表現と文書中で用いる表現とが異なるものとなっても、名詞句を構成しうる品詞の並びで構成された、より小さな単位の語句に分割して検索することにより検索漏れを減らすことができるという効果を得ることができる。

【００２９】図５は、名詞句が３つ以上の単語により構成されている場合の処理を示すフローチャートで、まず、ステップＳ１において、言語解析手段１２で抽出された名詞句が品詞情報つきで検索条件生成手段１３に入力される。ステップＳ２において、検索条件生成手段１３は、入力された名詞句が３つ以上の単語で構成されている場合は、ステップＳ３へと進む。ステップＳ３では、入力された名詞句の分割処理が行われる。例えば、入力された名詞句が“significant_JJ scientific_JJ discoveries_NN”（_JJは形容詞を表わす品詞タグ、_NNは名詞を表わす品詞タグである）という名詞句であった場合、これを“significant_JJ scientific_JJ,”“significant_JJ discoveries_NN,”“scientific_JJ discoveries_NN”という３つの単語ペアに分解する。

【００３０】次に、ステップＳ４において、単語ペアの選択処理が行われる。ステップＳ３で得られた単語ペアの各々について、例えば、以下に示すような予め設定しておいた単語ペア品詞構成条件との照合が行われる。その結果、単語ペア品詞構成条件に合致する単語ペアのみが選択される。上記の場合、条件に合致するものとして、“significant_JJ discoveries_NN”及び“scientific_JJ discoveries_NN”という２つの単語ペアが選択される。こうして、“significant_JJ scientific_JJ”のように２つの形容詞で構成される、名詞句を構成していない単語ペアは検索条件を生成するには不適切であるとして除外されることになる。なお、本実施例では、名詞句を構成しうる単語ペアを選択するようにしたが、表１に示す条件のもとに、名詞連続の単語ペアのみを選択するような構成とすることもできる。

【００３１】

【表１】

単語ペア品詞構成条件
NP → NN+NN
NP → JJ+NN
NP → CD+NN
・
・　　NP:名詞句　NN:名詞　JJ:形容詞　CD:基数

【００３２】次に、ステップＳ５において、言語解析手段１２で抽出された単語と検索条件生成手段１３により選択された単語ペアとを用いて検索条件の生成が行われ、検索条件生成処理を終了する。

【００３３】上述のように、本実施の形態では、名詞句が３つ以上の単語から構成される場合でも、名詞句を構成するような品詞の並びを持つ、２つの単語からなる単語ペアに分解して検索条件とするため、検索条件を適度に緩めることができ、検索漏れの可能性を低減して検索精度を向上させることができる。

【００３４】次に、本発明の他の実施形態、すなわち、言語解析手段で抽出された名詞句の中に他の句構造を持つ要素が含まれている場合について説明する。全体の機能構成は図２と同じであり、相違点は検索条件生成手段１３において、言語解析手段１２で抽出された名詞句の中に他の句構造を持つ要素が含まれている場合に、当該名詞句全体の句構造情報を基にして主要な語句を抽出して検索条件を生成するようにしたことにある。これは、

例えば、言語解析手段１２で抽出された名詞句が“foreign minorities in Germany”のように前置詞句を伴った複雑な句構造を有する場合でもこの中から主要な語句を抽出して検索条件を生成するようにしたことにより、検索漏れを減らすことができるという効果を得ることができる。

【００３５】図６は、名詞句の中に他の句構造を持つ要素が含まれている場合の処理を示すフローチャートで、まず、ステップＳ１１において、言語解析手段１２で抽出された名詞句が品詞情報つきで検索条件生成手段１３に入力される。ステップＳ１２において、検索条件生成手段１３は、入力された名詞句の中に他の句構造を持つ要素が含まれているかどうかを調べ、他の句構造を持つ要素が含まれている場合にはステップＳ１３へと進む。

【００３６】ステップＳ１３では、入力された名詞句全体の句構造情報を基にして主要な語句を抽出する処理が行われる。例えば、入力された名詞句が上記の“foreign minorities in Germany”である場合、言語解析手段１２によって同定された句構造情報は“NP [NP [foreign_JJ minorities_NN] PP [in_P NP [Germany_NN]]]”であるとする。この複雑な名詞句は、より小さな名詞句“foreign_JJ minorities_NN”とその名詞句を後置修飾している前置詞句“in_P NP [Germany_NN]”とで構成されていることがわかる。このとき、各句に含まれる、最小の名詞句の末尾の単語“minorities”と“Germany”が主要な語句として抽出される。末尾の単語を抽出するのは、英語では最小の名詞句においてその主要語（headと呼ばれる）は末尾に生起するという言語学的知見に基づくものである。

【００３７】次に、ステップＳ１４において、入力された名詞句の中に２つ以上の単語で構成されている最小名詞句があるかどうかを調べる。該当する名詞句がない場合には、ステップＳ１５に進み、単語及び主要語での検索条件生成処理が行われる。主要語を用いた検索条件の生成は、名詞句の場合と同様にＷＩＮＤＯＷ演算子を用いて表わされるが、２つの単語の間の距離や順序の指定は通常の名詞句の場合とは異なる表記が与えられる。これら２つの単語は共起しやすいという条件だけを課し、ある一定の距離内に離れて出現可能で、かつ出現順序は問わないという検索条件とした。

【００３８】一方、ステップＳ１４において、２つ以上の単語で構成される最小名詞句がある場合にはステップＳ１６へと進み、名詞句単位での検索条件生成処理が行われる。このとき、当該名詞句が３つ以上の単語で構成される場合には図５で説明したように名詞句分割処理を行って名詞句単位の検索条件を生成すればよい。

【００３９】こうして、最終的には、単語及び主要語での検索条件と最小名詞句での検索条件とがマージされることになる。例えば、上述の複雑な構造を有する名詞句に対して生成される、主要語での検索条件と最小名詞句での検索条件は以下のようになる。ここでは、主要語での検索条件生成に関して、単語が離れて出現可能な距離は、同一文中内とし、３０語と設定している。

主要語での検索条件：#scale [0.5] (#window [2, 30, u] (minorities, Germany))

最小名詞句での検索条件：#scale [0.5] (#window [1, 1, o] (foreign, minorities))

#window [2, 30, u] (minorities, Germany)は、“minorities”と“Germany”が任意の順序で２～３０語の範囲に出現する条件を指定している。

【００４０】上述のように、本実施の形態では、名詞句の中に他の句構造を持つ要素が含まれている場合でも、当該名詞句全体の句構造情報を基にして主要な語句を抽出して検索条件を生成するようにしたことにより、検索漏れを減らすことができるという効果を得ることができる。以上に説明した実施形態では、英文による文書を検索対象文書としたが、例えば、日本語、独語、仏語等の他の言語による文書でも本発明による文書検索を適用することもできる。

【００４１】

【発明の効果】以上に説明したように、本発明によれば、単語と名詞句に適切な重み付けを施して演算子により結合して検索条件が生成される。これにより、名詞句単位での検索条件と単語単位での検索条件とを合わせて検索条件を生成することができるので、検索漏れを低減し、検索精度を向上させることができる。

【００４２】また、本発明によれば、不要語リストを用意し、単語単位での検索条件を生成する場合と名詞句単位での検索条件を生成する場合とで不要語リストの使用条件を変えることにより、検索条件の要素となる適切な単語及び名詞句を検索要求文から抽出することができる。

【００４３】また、本発明によれば、３つ以上の単語で構成される名詞句の場合には、その構成単語の品詞情報を基にして検索条件の要素となる適切な２単語からなる単語ペアを選択することができる。このとき、３つ以上の単語で構成される名詞句を２つの単語で構成される単語ペアに分割して検索条件が生成されるので、検索条件の範囲を広げることができる。

【００４４】更に、本発明によれば、名詞句の中に他の句構造を持つ要素が含まれている場合でも、名詞句全体の句構造情報を基にして主要な語句を抽出して検索条件を生成するようにしたので、検索条件の範囲を広げることができる。

【図面の簡単な説明】

【図１】　本発明の第１の実施の形態による文書検索装置のブロック図である。

【図２】　図１に示した文書検索装置の機能ブロック図である。

【図３】　検索要求入力手段により検索要求文を入力し

11

たときの入力画面の一例である。

【図４】　検索結果の表示例を示す図である。

【図５】　名詞句が３つ以上の単語により構成されている場合の処理を示すフローチャートである。

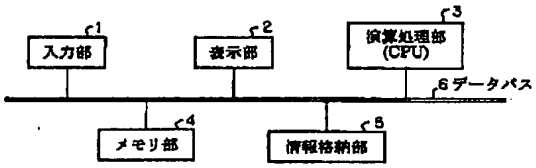
【図６】　名詞句の中に他の句構造を持つ要素が含まれている場合の処理を示すフローチャートである。

12

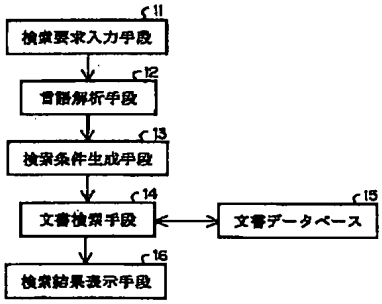
【符号の説明】

１…入力部、２…表示部、３…中央演算装置（ＣＰＵ）を含む演算処理部、４…メモリ部、５…情報格納部、６…データバス、１１…検索入力手段、１２…言語解析手段、１３…検索条件生成手段、１４…文書検索手段、１５…文書データベース、１６…検索結果表示手段。

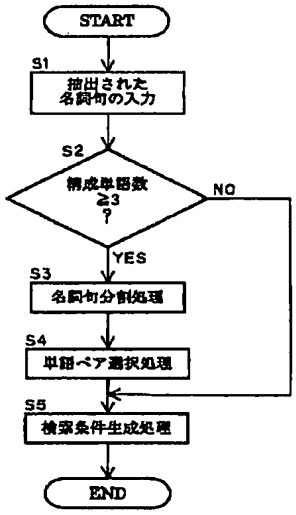
【図１】



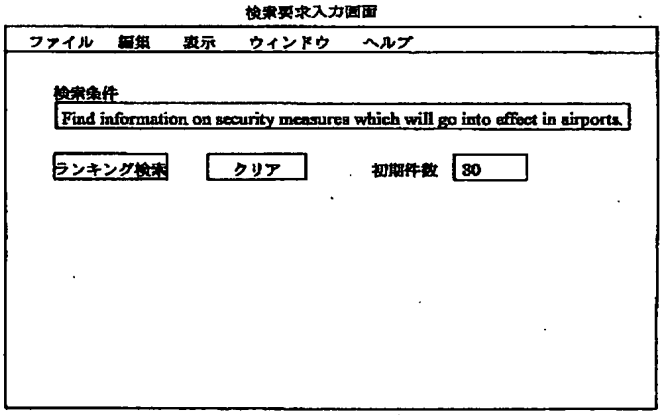
【図２】



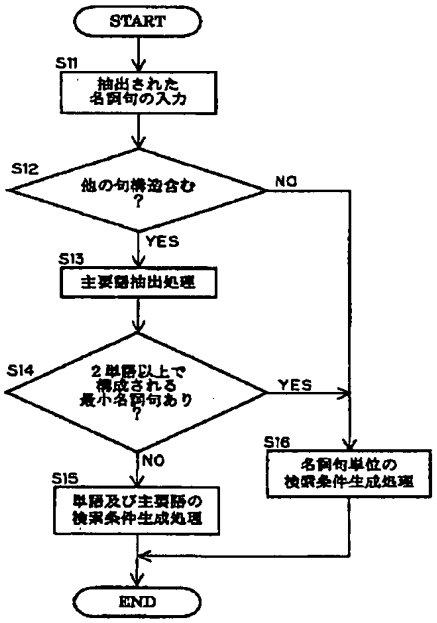
【図５】



【図３】



【図６】



【図４】

検索要求出力画面

ファイル　編集　表示　ウィンドウ　ヘルプ

検索条件

Find information on security measures which will go into effect in airports.

ランキング検索　クリア　初期件数 30

1 Long queues are commonplace at Beijing Capital International Airport's security gates now that all passengers flying to eight southeast coastal cities must .....

2 Foreign air carrier's security programs provide security procedures for foreign air carriers while operating to and from the United States, which is a .....

3 Ministers are this week expected to consider using army patrols to boost security at British airports after a third IRA mortar attack at Heathrow in five .....

4 A foreign tourist illegally carrying more than 120 grams of marijuana was arrested at the Shoudu Airport's security inspection station. This is the first .....